W176–W181 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm261

PathExpress: a web-based tool to identify relevant pathways in gene expression data

Nicolas Goffard and Georg Weiller*

ARC Centre of Excellence for Integrative Legume Research and Bioinformatics Laboratory, Genomic Interactions Group, Research School of Biological Sciences, Australian National University, GPO Box 475, Canberra, ACT 2601 Australia

Received January 30, 2007; Revised March 30, 2007; Accepted April 8, 2007

ABSTRACT

PathExpress is a web-based tool developed to interpret gene expression data obtained from microarray experiments by identifying the most relevant metabolic pathways associated with a subset of genes (e.g. differentially expressed genes). A graphical pathway representation permits the visualization of the expressed genes in a functional context. Based on the publicly accessible KEGG Ligand database, PathExpress can be adapted to any organism and is currently available for seven Affymetrix genome arrays. About 20% of the probe sets of each array have been assigned to Enzyme Commission numbers by homology relationship and linked to corresponding metabolic pathways.
PathExpress is available at http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/.

INTRODUCTION

Microarrays enable us to investigate the expression of thousands of genes simultaneously, providing a comprehensive overview of the gene activities in a given tissue. The results of such experiments are usually presented as lists of (differentially expressed) genes. A number of statistical tests have been employed for assessing differential gene expression (1) and several ontological tools are available (2–6) to support the biological interpretation of these data. Most are based on the identification of significant associations of gene ontology terms (7) with groups of genes, but this does not directly reflect metabolic networks. With the availability of biological pathway databases such as the kyoto encyclopedia of genes and genomes (KEGG) (8) or MetaCyc (9), several resources have been developed to visualize and analyse microarray data in the context of known biological networks (10–17). However, some limitations still remain, as current pathway databases are limited to organisms with well-annotated genomes and most of the analysis tools only attempt to match expression data to entire pathways without considering sub-pathways.

PathExpress overcomes some of these limitations, and provides a user-friendly web-based tool to interpret gene expression results from microarray experiments in the context of biological pathways. Based on the publicly available KEGG Ligand database of chemical compounds and reactions in biological pathways (18,19), PathExpress can be extended to any organism, as it uses similarity between the probe set sequences of supported genome arrays and the sequences of genes with known Enzyme Commission (EC) numbers in order to link probe sets to the metabolic networks. To take into account how reactions are linked in a pathway, sub-pathways are defined as a chain of reactions linked to each other by a common compound (substrate or product) (Figure 1). Two statistical approaches can be considered to perform a pathway analysis. The first compares a gene list to a pathway using a chi-squared test, a Fisher exact test or the hypergeometric distribution to calculate the probability of a specific number of genes from one pathway. The second is based on the analysis of all genes present on the genome array and measures the significance of pathway-level statistics computed from the gene-level statistics, using gene set enrichment analysis (20), random forests methods (21), Hotelling's T-square statistics (22) or random-set methods (23). With the aim of providing some flexibility to the user in defining his genes of interest, PathExpress compares a submitted list of genes to the genes involved in annotated pathways. The significantly overrepresented sets of reactions (pathways or sub-pathways) in the query list of genes are identified using a hypergeometric distribution test as developed in the BlastSets system (2). As the comparisons are based on enzyme compositions rather than single probe assignments, problems that arise from a multiplicity of genes coding for the same enzyme are largely overcome and the functional activities become apparent. In addition, an automatically generated graphical representation of the metabolic pathways

*To whom correspondence should be addressed. Tel: +61 2 6125 5916; Fax: +61 2 6125 7879; Email: Georg.Weiller@anu.edu.au

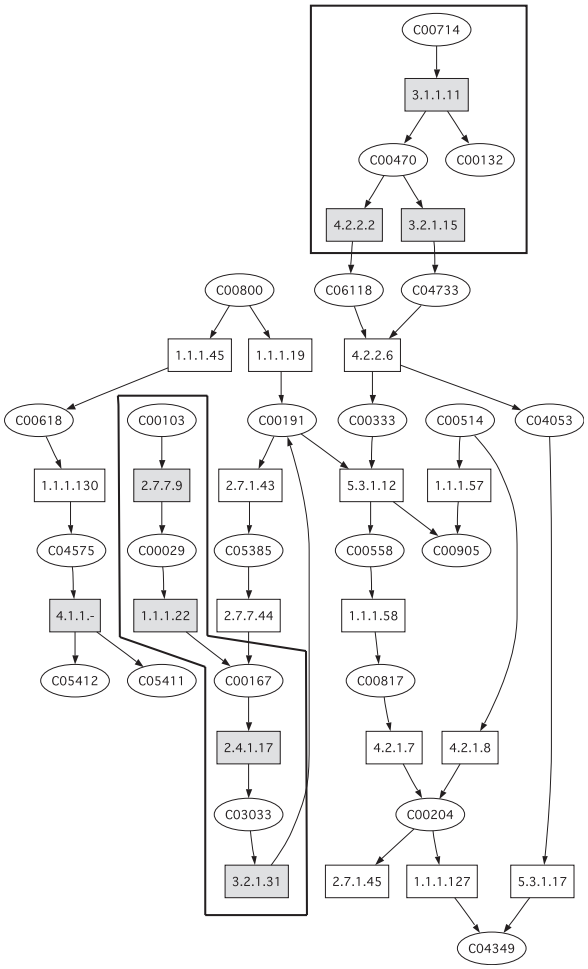

Figure 1. Representation of metabolic pathways and sub-pathways. The directed graph contains two types of nodes, compounds (labelled with their KEGG identifier and represented as ellipses) and reactions (labelled with the EC number of the enzyme involved and represented as boxes). Greyed reactions show that the corresponding enzyme has been identified in a given genome array. Directed arcs between two different nodes represent the consumption or the production of compounds by a reaction. The presented pathway contains eight enzymes with probe sets assignments. Two sub-pathways can be considered, containing three and four enzymes (surrounded by a black line).

allows the visualization of differential gene expression in a functional context.

PathExpress is freely available at http://bioinfoserver.rsbs.anu.edu.au/utils/PathExpress/.

METHODS

Data representation

PathExpress is based on a directed graph to model enzymatic reactions in the context of biological pathways (Figure 1). Two types of nodes are used to represent compounds and reactions that can be mediated by one or more enzymes. Directed edges, connecting these different nodes, correspond to the consumption or the production of compounds by the reactions. The data used to build this network is derived from the Compound, Reaction and Enzyme sections of the publicly available KEGG Ligand database (18,19).

To link gene expression data to pathways, PathExpress uses pre-computed assignments of the probe sets of supported genome arrays to EC numbers, identifying enzyme entries. These assignments are based on sequence similarities with proteins retrieved from the Swiss-Prot database (24). Blastx (25) is used to find the best match (E-value $\leq 10^{-8}$) for the sequences representing each probe set sequence (i.e. sequences derived from the most 5′ to the most 3′ probe in the public UniGene cluster) of the genome array analysed. If these entries have been annotated as an enzyme, the probe set is assigned to the corresponding EC number, extracted from its definition line. Note that probe sets that cannot be assigned to EC numbers are excluded from further analyses, and although this limits the number of usable probe sets, it also eliminates much of the ambiguity that arises from multiple (iso) genes encoding the same enzymatic function. This strategy can be applied to any set of sequences.

As of March 2007, data from eight Affymetrix Genome Arrays are available in PathExpress (Table 1). They were selected because of their importance as model organisms for various taxonomic groups or their economic interest, even if they don't have a well–annotated genome. Additional species could be included upon request.

Microarray data analysis

To interpret gene expression results from microarray experiments, PathExpress detects if the genes associated with a pathway or sub-pathway are statistically over-represented in a set of sequences, when compared to the rest of the genome array. When a list of identifiers has been submitted, PathExpress first assigns them to EC numbers according to pre-computed relationships. The proportion of submitted EC numbers is then tested for every (sub) pathway. For each test, a *P*-value, representing the probability that the intersection of the given list with the list of enzymes belonging to the given set of reactions occurs by chance, is calculated using the hypergeometric distribution (26). Because multiple hypothesis tests are performed, it is necessary to correct these *P*-values. Two adjustment methods are available in PathExpress; the conservative Bonferroni correction method (27) in which the *P*-values are multiplied by the number of comparisons and the less stringent False Discovery Rate (FDR) approach (28) defined as the determination of the expected proportion of false positive results among all rejected hypotheses.

SYSTEM IMPLEMENTATION

The PathExpress web server runs on a Linux server (2 Intel 4 3.20 GHz, 1 GB RAM). It combines a PostgreSQL database management system to store the data with a dynamic web interface based on PHP and Perl. Data pre-processing is implemented in Perl, statistical analyses are performed using Perl and the R statistical package and graphical representations are generated using GraphViz software (http://www.graphviz.org).

Table 1. Available Affymetrix genome arrays and assignment statistics

Affymetrix Genome Array	Organism	Sequences[a]	Assigned sequences[b]	EC[c]
ATH1 Genome Array	*Arabidopsis thaliana*	22 765	5 177	823
Drosophila Genome 2.0 Array	*Drosophila melanogaster*	18 952	3 107	724
E. coli Genome 2.0 Array	*Escherichia coli*	10 208	2 245	803
Human Genome U133 Plus 2.0 Array	*Homo sapiens*	39 070	3 332	658
Medicago Genome Array	*Medicago truncatula*	50 900	8 981	953
Rice Genome Array	*Oryza sativa*	57 194	10 068	923
Soybean Genome Array	*Glycine max*	37 618	6 502	803
Yeast Genome 2.0 Array	*Saccharomyces cerevisiae*	5 814	1 471	601
Yeast Genome 2.0 Array	*Schizosaccharomyces pombe*	5 028	1 333	566

[a]Number of probe set sequences.
[b]Number of probe set sequences assigned to an EC number.
[c]Number of distinct EC numbers corresponding to the probe set sequences.

An analysis of 1000 identifiers (i.e. comparison to sub-pathways with FDR adjustments) takes less than 2 s and the automatic generation of a graphical representation takes less than 1 s.

WEB INTERFACE

Input

The input data for PathExpress consists of a list of genes of interest (Affymetrix probe set identifiers and/or GenBank accession numbers) present in the selected genome array. Other parameters can be specified: the type of comparison (pathway or sub-pathway), the *P*-value significant threshold and the adjustment method for multiple testing.

Output

The PathExpress output contains the list of pathways or sub-pathways that are significantly associated with the enzymes in a list of submitted sequence identifiers (Figure 2a). Metabolic pathways are ranked by increasing *P*-values whereas sub-pathways are grouped according to the pathway to which they belong. In each case, those that are significant (according to the *P*-value threshold defined by the user) are highlighted. Each pathway can be displayed as an automatically generated graphical representation (Figure 2b) and as an enumeration of reactions. On these pictures, reactions are highlighted if the according enzyme was identified in the genome array (in grey) and in the submitted list of identifiers (in yellow). The name of the compounds and the definition of the reactions are displayed as a tool-tip when the mouse is over any of the nodes in the graph. In addition, compounds are linked on the corresponding KEGG entry. If the user clicks on a reaction node, a new page containing the description of the enzymes associated with the list of probe sets assigned in the selected genome array is opened (Figure 2c). Blast results used for the EC assignments are available for each probe set in its 'detail' page.

All results can be downloaded as tab-delimited text files for further statistical analyses. Pictures representing the pathways can be saved in png or dot format and visualized locally using the GraphViz software. To enhance the visualization of the expression of individual probe sets, all resources (EC assignments and pictures with xml descriptions) are available to be imported into MapMan (17). Although initially developed for the *Arabidopsis thaliana* array, MapMan has been extended to other organisms, and classification including the Affymetrix arrays of the plants is presented in Table 1 (29).

COMPARISON WITH EXISTING TOOLS

To illustrate the novelty and utilities of PathExpress, we compare it with existing web-based pathway analysis tools (Table 2). Except for the KOBAS server, these tools are limited to organisms with well-annotated genomes. Indeed, the KOBAS server annotates a set of genes with KEGG Orthologous (KO) terms to link them to KEGG pathways (13). To overcome the problems that arise from a multiplicity of genes coding for the same enzyme and thus to provide a robust pathway identification, PathExpress focuses on chemical compounds and reactions in the KEGG Ligand database. It uses pre-computed assignments of probe set sequences to EC numbers and thus can be extended to any organism or set of sequences. Another limitation of these tools is that they define a metabolic pathway as a set of genes without considering the relationships between the reactions within a pathway. To take into account these relationships, we defined a sub-pathway as a chain of enzymatic reactions linked to each other by a common compound (substrate or product) within a pathway (Figure 1). This strategy allows us to identify the most relevant sets of reactions (pathways or sub-pathways) associated with a subset of genes.

CONCLUSION

We have developed PathExpress, a web-based tool for finding the pathways relevantly affected in gene expression experiments. The focus on enzymes results in robust pathway identification. PathExpress can also correctly identify partial pathways, and provides a graphical representation for their visualization. Based on the

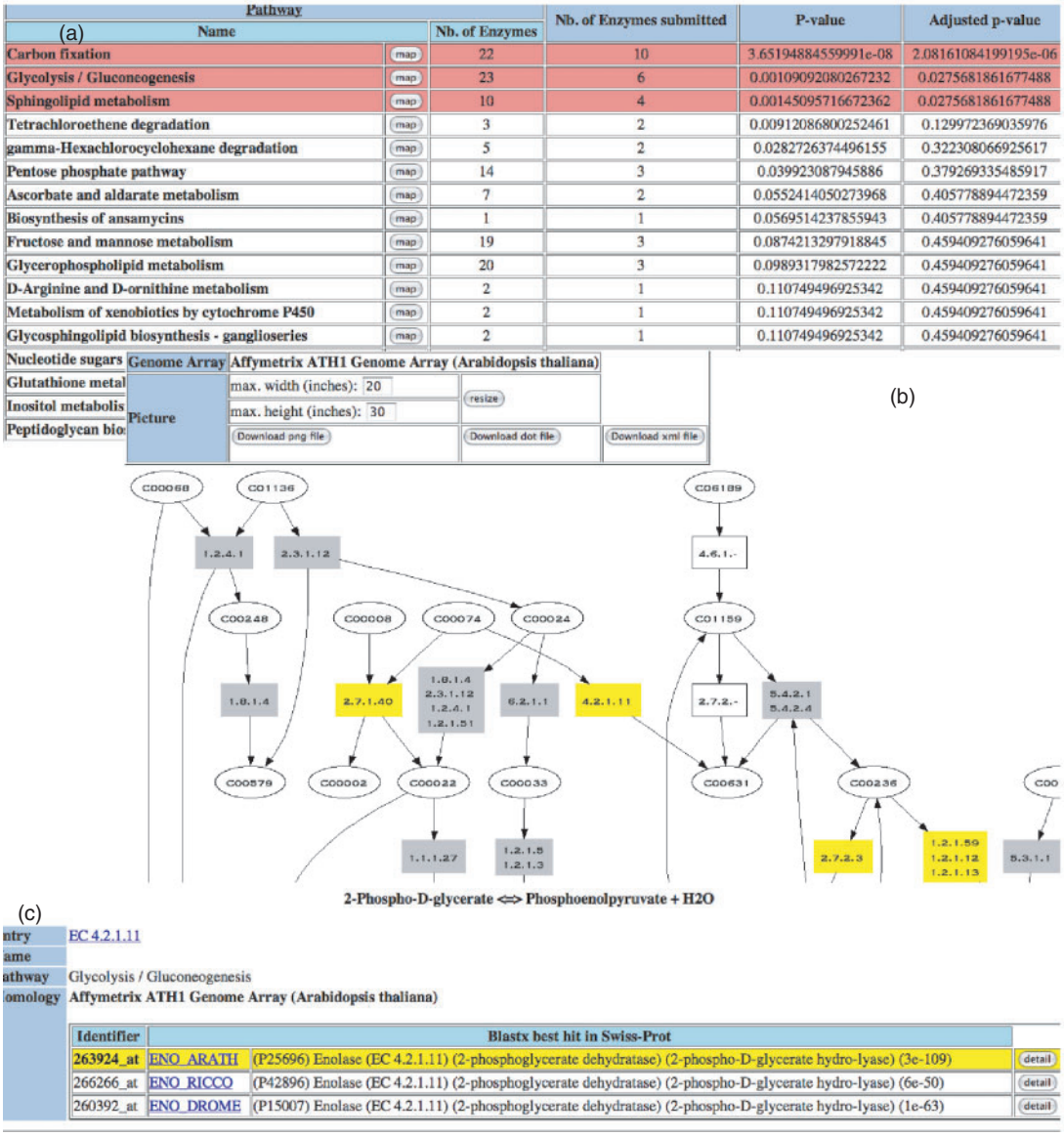

Figure 2. Screenshots of the PathExpress web interface. (**a**) The list of metabolic pathways whose enzyme composition intersects with the enzymes corresponding to a list of submitted identifiers, ordered by increasing *P*-value. Each row reports information concerning the pathway's name and the number of enzymes. The comparison of groups is reported with the number of submitted enzymes involved in the pathway and the *P*-value for finding the group by chance, associated with the corresponding adjusted *P*-value. The significant pathways are highlighted in red ('carbon fixation', 'glycolysis/gluconeogenesis' and 'pentose phosphate pathway' in this example). (**b**) Graphical representation of the glycolysis/gluconeogenesis pathway for enzymes identified in the Affymetrix ATH1 Genome Array (*Arabidopsis thaliana*). Reactions mediated by enzymes found in the genome array are highlighted in grey whereas those where the enzymes are also present in the query are highlighted in yellow. (**c**) Example of the detail page of an enzymatic reaction. Each enzyme is reported with its EC number linked to the KEGG database (Entry), its recommended and alternative names, the pathways in which this enzyme is involved (glycolysis/gluconeogenesis in this example) and the list of probe sets assigned. For each probe set, the identifier and description of the best match in Swiss-Prot is displayed. The 'detail' button is linked to the complete blast report. The corresponding row is highlighted in yellow if the probe set belongs to the submitted list of identifiers.

KEGG Ligand database and on a pre-computed assignment of probe sets to EC numbers, PathExpress can be extended to any organism or set of analysis sequences (e.g. custom DNA microarray, proteome array) and hence provides a useful resource for the integration of transcriptomic and proteomic data sets.

In the near future, additional species will be included in PathExpress. The process of assigning probe sets to EC numbers will be improved by taking into account the domain composition of the protein and by using enzyme-specific profiles. We also consider applying other pathway-based tests to analyse all genes from expression data. Finally, we intend to further develop the system by extending its application to the analysis of metabolomic results, since compound information is already included in the metabolic network representation.

Table 2. Comparison of PathExpress with existing tools

Software	Pathways	Input	Output	Comments	References
PathExpress	KEGG	List of identifiers (Affymetrix identifiers and/or Gene accession numbers)	Statistically significant pathways and sub-pathways with graphical visualisation	Affymetrix data but extendable to any organism or set of sequences	
ArrayXPath II	GenMAPP, KEGG, BioCarta, PharmGKB	Clustered gene expression profile	Statistically significant pathways with graphical visualisation	*Homo sapiens*, *Mus musculus*, *Rattus norvegicus*	(10)
KOBAS	KEGG	List of KO identifiers	Statistically significant pathways	Annotation of a set of genes or proteins with KO terms	(13)
PathMAPA	KEGG, TAIR, NCBI, GO	Expression data file (GenePix, Affymetrix)	Statistically significant pathways, enzymes, genes with graphical visualisation	*Arabidopsis thaliana*	(14)
PathwayExplorer	KEGG, BioCarta, GenMAPP	Expression data file (Tab-delimitated)	Statistically significant pathways with graphical visualisation	Well-annotated genomes	(15)
Pathway Miner	GenMAPP, KEGG, BioCarta	List of Gene Accession numbers with gene expression values	Statistically significant pathways with graphical visualisation	*Homo sapiens*, *Mus musculus*	(16)

ACKNOWLEDGEMENTS

This study was funded by an Australian Research Council Centre of Excellence grant. Funding to pay the Open Access publication charges for this article was provided by the same grant.

Conflict of interest statement. None declared.

REFERENCES

1. Cui,X. and Churchill,G.A. (2003) Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol.*, **4**, 210.
2. Barriot,R., Poix,J., Groppi,A., Barre,A., Goffard,N., Sherman,D., Dutour,I. and de Daruvar,A. (2004) New strategy for the representation and the integration of biomolecular knowledge at a cellular scale. *Nucleic Acids Res.*, **32**, 3581–3589.
3. Goffard,N. and Weiller,G. (2007) GeneBins: a database for classifying gene expression data, with application to plant genome arrays. *BMC Bioinformatics*, **8**, 87.
4. Cheng,J., Sun,S., Tracy,A., Hubbell,E., Morris,J., Valmeekam,V., Kimbrough,A., Cline,M.S., Liu,G. *et al.* (2004) NetAffx Gene Ontology Mining Tool: a visual approach for microarray data analysis. *Bioinformatics*, **20**, 1462–1463.
5. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res.*, **31**, 3775–3781.
6. Al-Shahrour,F., Diaz-Uriarte,R. and Dopazo,J. (2004) FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, **20**, 578–580.
7. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
8. Kanehisa,M. and Goto,S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
9. Caspi,R., Foerster,H., Fulcher,C.A., Hopkinson,R., Ingraham,J., Kaipa,P., Krummenacker,M., Paley,S., Pick,J. *et al.* (2006) MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res.*, **34**, D511–D516.
10. Chung,H.J., Park,C.H., Han,M.R., Lee,S., Ohn,J.H., Kim,J., Kim,J. and Kim,J.H. (2005) ArrayXPath II: mapping and visualizing microarray gene-expression data with biomedical ontologies and integrated biological pathway resources using Scalable Vector Graphics. *Nucleic Acids Res.*, **33**, W621–W626.
11. Baitaluk,M., Sedova,M., Ray,A. and Gupta,A. (2006) BiologicalNetworks: visualization and analysis tool for systems biology. *Nucleic Acids Res.*, **34**, W466–W471.
12. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nat. Genet.*, **31**, 19–20.
13. Wu,J., Mao,X., Cai,T., Luo,J. and Wei,L. (2006) KOBAS server: a web-based platform for automated annotation and pathway identification. *Nucleic Acids Res.*, **34**, W720–W724.
14. Pan,D., Sun,N., Cheung,K., Guan,Z., Ma,L., Holford,M., Deng,X. and Zhao,H. (2003) PathMAPA: a tool for displaying gene expression and performing statistical tests on metabolic pathways at multiple levels for Arabidopsis. *BMC Bioinformatics*, **4**, 56.
15. Mlecnik,B., Scheideler,M., Hackl,H., Hartler,J., Sanchez-Cabo,F. and Trajanoski,Z. (2005) PathwayExplorer: web service for visualizing high-throughput expression data on biological pathways. *Nucleic Acids Res.*, **33**, W633–W637.
16. Pandey,R., Guru,R.K. and Mount,D.W. (2004) Pathway Miner: extracting gene association networks from molecular pathways for predicting the biological significance of gene expression microarray data. *Bioinformatics*, **20**, 2156–2158.
17. Thimm,O., Blasing,O., Gibon,Y., Nagel,A., Meyer,S., Kruger,P., Selbig,J., Muller,L., Rhee,S. *et al.* (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J.*, **37**, 914–939.
18. Goto,S., Nishioka,T. and Kanehisa,M. (1998) LIGAND: chemical database for enzyme reactions. *Bioinformatics*, **14**, 591–599.
19. Goto,S., Okuno,Y., Hattori,M., Nishioka,T. and Kanehisa,M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res.*, **30**, 402–404.
20. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
21. Pang,H., Lin,A., Holford,M., Enerson,B.E., Lu,B., Lawton,M.P., Floyd,E. and Zhao,H. (2006) Pathway analysis using random forests classification and regression. *Bioinformatics*, **22**, 2028–2036.
22. Kong,S.W., Pu,W.T. and Park,P.J. (2006) A multivariate approach for integrating genome-wide expression data and biological knowledge. *Bioinformatics*, **22**, 2373–2380.
23. Newton,M.A., Quintana,F.A., den Boon,J.A., Sengupta,S. and Ahlquist,P. (2007) Random-set methods identify distinct aspect of the enrichment signal in gene-set analysis. *Ann Appl Stat.*, in press.
24. Bairoch,A., Apweiler,R., Wu,C., Barker,W., Boeckmann,B., Ferro,S., Gasteiger,E., Huang,H., Lopez,R. *et al.* (2005)

The Universal Protein Resource (UniProt). *Nucleic Acids Res*., **33**, D154–D159.
25. Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol*., **215**, 403–410.
26. Cho,R.J., Huang,M., Campbell,M.J., Dong,H., Steinmetz,L., Sapinoso,L., Hampton,G., Elledge,S.J., Davis,R.W. *et al.* (2001) Transcriptional regulation and function during the human cell cycle. *Nat. Genet*., **27**, 48–54.
27. Bonferroni,C.E. (1936) Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del Regio Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.
28. Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B*, **57**, 289–300.
29. Goffard,N. and Weiller,G. (2006) Extending MapMan: application to legume genome arrays. *Bioinformatics*, **22**, 2958–2959.